How to Improve TTS Systems for Emotional Expressivity

Antonio Rui Ferreira Rebordao, Mostafa Al Masum Shaikh, Keikichi Hirose and Nobuaki Minematsu

Department of Information and Communication Engineering, University of Tokyo, Japan
{antonio, almasum, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Several experiments have been carried out that revealed weaknesses of the current Text-To-Speech (TTS) systems in their emotional expressivity. Although some TTS systems allow XML-based representations of prosodic and/or phonetic variables, few publications considered, as a pre-processing stage, the use of intelligent text processing to detect affective information that can be used to tailor the parameters needed for emotional expressivity. This paper describes a technique for an automatic prosodic parameterization based on affective clues. This technique recognizes the affective information conveyed in a text and, accordingly to its emotional connotation, assigns appropriate pitch accents and other prosodic parameters by XML-tagging. This pre-processing assists the TTS system to generate synthesized speech that contains emotional clues. The experimental results are encouraging and suggest the possibility of suitable emotional expressivity in speech synthesis.

Index Terms: emotional expressivity, speech synthesis, TTS, MaryXML, intelligent text processing, affect sensing

1. Motivation

Expressivity and emotional eloquence are relevant issues to improve the perception of synthesized speech [1, 2, 3, 4, 5]. Appropriate tone, pitch accent and suitable speech intensity can help conveying speech subtleties in a contextual and content-rich manner and TTS systems should generate speech that sounds as natural as human speech. However, contemporary TTS systems tend to synthesize text in a way that sounds unnatural due to deficiencies in the syntactic analysis of the input text. However, this problem can be solved by an efficient extraction of affective clues that could be used during the synthesis phase. This is our main motivation and we conducted several experiments that assess the emotional expressivity of synthesized and human speech samples. The figure 1 shows the relative changes of four quantitative speech variables namely, Speech Rate (SR) (i.e., syllable/sec), Pitch Average (PA), Pitch Range (PR) and Intensity (I) with respect to neutral human speech. This evaluation results match the findings of the studies [1, 6] and ideally a TTS system should also match this behavior.

In our experiments we considered the latest versions (March 2009) of six TTS systems (Loquendo, RealSpeak, AT&T, MacPlainTalk, Festival and Mary TTS) but due to space limitation we only present (in figure 2 and 3) the evaluation results of the three best performing TTS systems. According to these results it is concluded that TTS systems fail to incorporate emotion subtleties. For example, to signal sadness in the synthesized speech the SR and PA should be slightly slower; PR should be slightly narrower; and the Intensity of the signal should be lower compared with neutral speech. On the contrary, to signal happiness the SR should be faster or slower; the PA should be much higher; the PR should be much wider; and the Intensity should be higher. However, it can be observed in figure 4 that some TTS systems do not necessarily perform in this way.

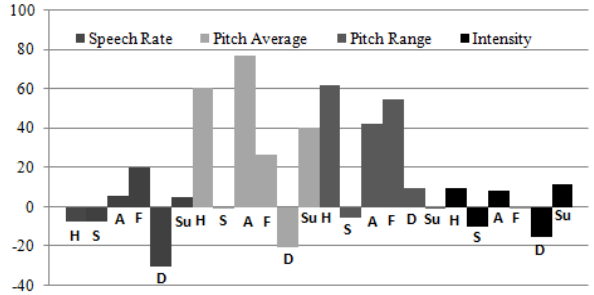

Figure 1: *Changes (in percentage) of SR, PA, PR, I for 6 emotions (from left to right and per color: happy (H), sad (S), anger (A), fear (F), disgust (D) and surprise (Su)) for speech articulated by humans (with respect to neutral speech).*

Figure 2 shows that our subjects failed to perceive emotions in synthesized speech samples and few people could perceive the right emotion (as it can be observed in figure 3). This suggests that TTS systems fail in conveying emotional expressiveness and that people usually can not perceive the emotion and subtleties that are associated with the speech content. However, these features sometimes are vital to lucid understanding of the speech intention and meaning.

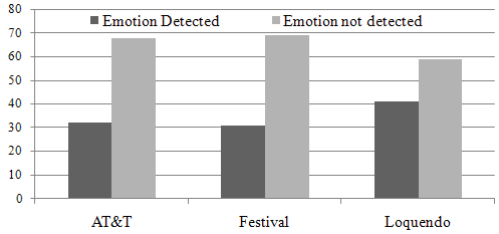

Figure 2: *Results (in percentage) of the perceptual test regarding the emotion recognition of the speech samples synthesized by AT&T, Festival and Loquendo.*

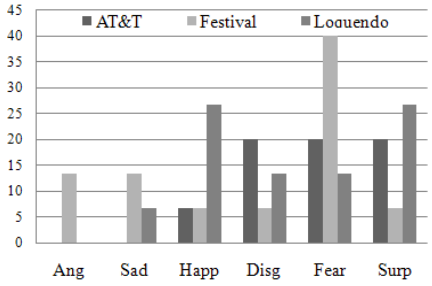

Figure 3: *Results (in percentage) of the perceptual test regarding the efficiency of the emotion recognition (six emotions) of the speech samples synthesized by AT&T, Festival and Loquendo.*

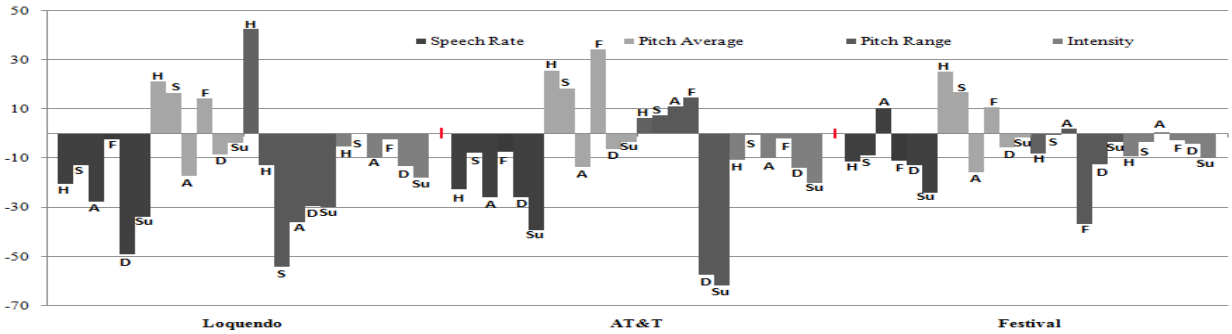

Figure 4: *Changes (in percentage) of SR, PA, PR and I with respect to neutral speech for the TTS systems Loquendo, AT&T and Festival. Each bar corresponds to an emotion (from the left to the right and per color: happy (H), sad (S), anger (A), fear (F), disgust (D) and surprise (Su).*

2. Related Work

Although tremendous effort has gone into speech synthesis and affective automatic assessment of speech, as far as we know, there is no system that takes the content (e.g., typed text), evaluates its affective information and parameterizes appropriate prosodic settings that can feed a TTS engine. By reviewing carefully the existing literature it is found that research regarding expressivity in synthetic speech is closely related to the following concepts: emotional text-to-speech synthesis; control languages to guide TTS synthesis process; flexibility in TTS architecture; and emotion recognition from textual data. The following sections briefly discuss these concepts.

2.1. Emotional Speech Synthesis

Previous researches (e.g., [2, 3, 4, 5, 6]) have found that there are several features in human speech that are related with its affective content. These features refer to: different statistical values (e.g., max, mean, standard deviation, etc.) of the fundamental frequency F0; different statistical values of the first three formants (F1, F2, and F2); and their bandwidths (BW1, BW2, and BW3), energy, speaking rate, etc. Generally these features are derived by observing how human's voice changes accordingly to different emotions. The studies mentioned above have established that when a speaker is in a state of fear, anger or joy, then his speech is typically faster, louder, and enunciated, with strong high-frequency energy. When the speaker is bored or sad, then his speech is typically slower and low-pitched, with very little high-frequency energy. Such pragmatic knowledge obtained from speech signal processing has inspired various kinds of synthesis methods like, formant synthesis, diphone concatenation, unit selection and prosody rules based synthesis. In [3, 4] these techniques are described along with their advantages and disadvantages. Moreover, techniques like explicit prosody control [1, 5, 7], expressivity based unit selection [8], HMM based parametric synthesis [9], non-verbal vocalization [10], etc., are quite popular and obtained partial success for recognizing anger and sadness in synthesized speech samples.

2.2. Sensing Affective Information from Text

This research addresses the aspect of subjective opinion, particularly the identification of different emotive dimensions and the classification of texts by their emotion affinity. It can be argued that the affective content of a text and its analysis depend on the audience, context and world knowledge. The assessment of affective information from text is based in one, or in a combination of the following techniques: keyword spotting; lexical affinity; statistical methods; a dictionary of affective concepts and lexicon; common-sense knowledge-base; fuzzy logic; knowledge-base from facial expression; machine learning; domain specific classification and contextual valence assignment. Some researches [16, 17, 18, 19] dealt with the above techniques. For example, Shaikh et al. [18], implemented a technique based on contextual valence assignment and achieved tremendous results in recognizing different emotions (e.g. happiness, sadness, anger, etc.) from text and, Liu et al. [19], using common-sense knowledge could detect the six basic emotions in a text.

2.3. MARY TTS: A Flexible TTS System

The MARY TTS system [20] is a client-server application written in Java and created at DFKI GmbH. MaryXML serves as the configuration input language of this system and thus has become a very flexible toolkit for speech synthesis research. We have chosen MARY TTS system because it allows the dynamic creation of MaryXML with appropriate prosodic and accent properties that relate with the intended emotion and allows us access to all intermediate processing results for the purposes of debugging and analysis.

3. Our Approach

Our system deals with six basic emotions: happy, sad, fear, anger, surprise and disgust. It performs affective evaluation of the input text and, accordingly to the emotional content of the input sentence, produces MaryXML that matches the desired prosodic parameters and the findings reported in [1, 2, 4, 6, 5]. This Dynamic MaryXML is used as input for MARY TTS system and assists the speech synthesis process.

3.1. System Architecture

A pipeline architecture with the following steps is followed: Language Processing, Textual Affect Sensing and Generating Dynamic MaryXML. These steps are briefly described as following:

3.1.1. Language Processing

For each input sentence the language processing module outputs triplet(s) consisting of a subject or agent, a verb and an object. Each member of the triplet may or may not have associated attribute(s) (e.g., adjective, adverb, etc.). A XML-formatted syntactic and functional dependency information for each word of the input text is obtained using the

Machinese Syntax parser [21] and this output constitutes the basis for further processing that generates the triplet(s). Since a triplet is initiated with an occurrence of a verb in the sentence, the semantic parser may obtain more than one such triplet if there are multiple verbs in the sentence. Basically a triplet encodes information about "who is associated with what and how" with a notion of semantic verb frame analysis. For example, the sentence *"The car exploded near a popular ice cream parlor, sending flames and shrapnel through the busy square and killing 17 people."* produces three triplets as shown in Table 1.

Table 1: *Triplet output of the parser for the above example.*

Triplets processed by the Semantic Parser	
Triplet 1	[[['*Actor:'*, 'car', *'Actor-Type:'*, 'object', *'Actor-Attrib:'*, ['DET: the']], ['*Action-Name:'*, 'explode', *'Action Status:'*, 'Past', *'Action-Attrib:'*, ['place: near a popular ice cream parlor']], ['*Object-Name:'*, '', *'Object-Type:'*, '', *'Object-Attrib:'*, ['']]]
Triplet 2	[[['*Actor:'*, '', *'Actor-Type:'*, '', *'Actor-Attrib:'*, []], ['*Action-Name:'*, 'send', *'Action-Status:'*, 'Present Progressive', *'Action-Attrib:'*, ['place: through the busy square']], ['*Object-Name:'*, 'flame and shrapnel', *'Object-Type:'*, 'N NOM', *'Object-Attrib:'*, ['']]]
Triplet 3	[[['*Actor:'*, '', *'Actor-Type:'*, '', *'Actor-Attrib:'*, []], ['*Action-Name:'*, 'kill', *'Action-Status:'*, 'Present Progressive', *'Action-Attrib:'*, []], ['*Object-Name:'*, 'people', *'Object-Type:'*, 'N NOM', *'Object-Attrib:'*, ['Quantity: 17']]]]

3.1.2. Textual Affect Sensing

We used the output of the system SenseNet developed by Shaikh et al. [18] that can process the triplet formatted input of a sentence. SenseNet can perform affective sentence sensing by assessing the contextual valence of the words using rules and prior-valence values of the words. It outputs a numerical value ranging from -15 to +15 flagged as the "sentence-valence" for each sentence that is used as input. For example, SenseNet outputs -10.76 for the sentence referred above as an example. The output value indicates a numerical measure of negative or positive sentiments carried by the sentence. SenseNet implements a cognitive theory of emotion known as the OCC emotion model [22] by developing rules for the model defined emotions. Therefore it can classify input texts according to eight types of emotions, namely, happy, sad, hope, fear, admiration, shame, love and hate, plus a neutral category. In this system, the output of SenseNet is mapped to the basic six emotions in the following manner: happy, hope and love are considered as happiness, sad as sadness, fear as fear, admiration as surprise, shame as anger and hate as disgust. Following an experimental study [18], the accuracy of SenseNet to assess sentence-level negative/positive sentiment is 91% and classification accuracy of eight emotion types is 82%.

3.1.3. Dynamic MaryXML Generation

After the input text has been processed as mentioned above, we obtain the affective assessment of the text: the overall emotion carried by the text; the positive or negative meaning of the events represented by the triplet(s); and the attributes (e.g., location, time, etc.) of the events that are considered important. First, several speech parameters are set for the overall negative or positive affective connotation of the text and then parameters like pitch, pitch-dynamics, number-of-pauses, etc., are adjusted accordingly to the detected emotions. For example, if a sentence would have to express "happiness", then the overall speech rate is set faster, pitch average is set higher, pitch range is set much wider, intensity is made higher, and pitch changes are set as smooth upward. The phrasal tones (L-L%, L-H%, H-H%, and H-L%) and the pitch accents (peak, low, scooped, and rising peak) are considered at word and phrase level and are assigned using ToBI notation.

The MaxyXML offers a rich set of prosody attributes that allow parameterization that suits the desired emotion. Currently the MARY TTS system has the following natural language components: Tokenize; Preprocessing; and Tagger & Chunker. These components can process an input text or sentence given in MaryXML format but our system, at present, has nothing to do with these components. Our system, from plain text, creates prosody-rich MaryXML that can be processed by the MARY TTS system and perform the synthesis process, mainly in an affective context. In future we plan to add a pre-processing module to the MARY system that implements our approach by performing emotion recognition from the plain text and automatic generation of MaryXML that matches the parameters to convey the emotion recognized in the text.

3.2. Example Output

The following is an example of the dynamic MaryXML for the sentence (related with fear) that we used in Table 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML" version="0.4"
xml:lang="en">
<prosody pitch="-5%" pitch-dynamics="-25%"
range="5.32st" range-dynamics="+26%" preferred-accent-
shape="falling" accent-slope="+75%" accent-
prominence="+58%" preferred-boundary-type="low" rate="-
0%" number-of-pauses="+23%" pause-duration="-7%"
vowel-duration="-5%" nasal-duration="-5%" liquid-
duration="-5%" plosive-duration="+41%" fricative-
duration="+41%" volume="61">
The car exploded near a popular ice cream parlor, sending
flames and shrapnel through the busy square and killing 17
people.
</prosody>
</maryxml>
```

4. Experiments and Results

For each affective text, using MARY TTS system, we created two versions of synthesized speech samples (thus a total of 40 speech samples). One is the output obtained by just using the plain text and the other is obtained from the dynamic MaryXML outputted by our approach. Both cases used the voice Mbrola-us2, version 3.5.0, and the length of each synthesized speech audio sample is 17 seconds on average. Therefore, we have two systems, the plain text input system (S1) and the dynamic MaryXML input system (S2). The online survey that was conducted through the link, http://research.rebordao.ne/emostory/, had a total of 15 participants (all of them were non-English speaking natives). The subjects had to listen to the synthesized speech audio

samples produced by S1 and S2. They were asked to assess if they could perceive any emotion, or not. If an emotion would be perceived, it would be asked them to select one of the 6 basic emotions. We considered the scores obtained from the web-survey for which, either one or both systems, received an emotion perception score. Figure 5 shows that the subjects perceived easily the emotion from the real audio (excepting for disgust). Furthermore, the system S2 performed significantly better than S1 for conveying anger (improvement of 14.3%), disgust (improvement of 30.4%) and happiness (improvement of 28.6%) but for conveying sadness, S1 performs better (40% for S2 and 75% for S1). This could occur due to the tendency that S1 produces synthesized speech with intonational information related to negative emotions and therefore, the subjects usually perceive all the output of S1 as sad.

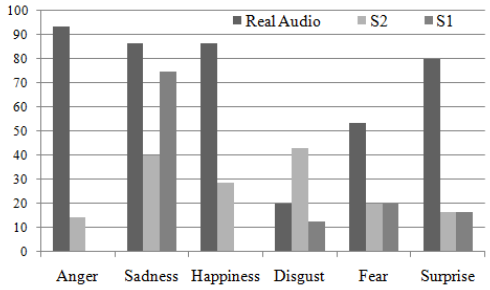

Figure 5: *The emotion recognition efficiency rates (in percentage) of the perceptual test for the Real Audio, S2 and S1.*

The results are encouraging in two manners, firstly S1 is very weak to convey positive emotion (e.g., "happiness") and our approach can solve this problem. Secondly, S1 has a tendency to express negative emotions (e.g., "sadness") and our approach can also be applied at this level to incorporate different levels of negativity/positivity for the individual phrases of a sentence. These help a better emotion perception of the synthesized speech.

5. Conclusion

There are numerous research works and techniques that aim at incorporating expressiveness in synthesized speech and this can be achieved by creating speech that conveys suitable emotions. In our study we have found that several well-known TTS systems, particularly MARY TTS system, do not produce affective synthesized speech. However, this situation can be improved by pre-processing the input in two manners, first by recognizing the emotions conveyed through the plain text and then controlling the synthesis process by assigning appropriate prosodic parameters that suit the detected emotions. Thus, the output of our system is an enhanced XML-based (i.e., dynamic MaryXML) interpretation of the plain input text that is given to the TTS system (i.e., MARY TTS) to process. A perceptual test was performed using the synthesized speech produced from the enriched XML-based input and the results support that these speech samples are more affectively expressive than the speech samples synthesized from the plain text. As future work we plan to build a tool combining all the resources discussed in our approach and add it as an add-on to MARY TTS system. Among its possible applications, it could allow speech impaired people to generate synthesized speech that conveys appropriate emotions just by typing text into a computer or into other devices.

6. References

[1] Cahn, J.E., The generation of affect in synthesized speech. Journal of the American Voice I/O Society, 8, 1–19, 1990.

[2] Oudeyer, P., The production and recognition of emotions in speech: features and algorithms. International Journal of Human—Computer Studies, 59, pp. 157–183, 2003.

[3] Schröder, M., Expressive Speech Synthesis: Past, Present, and Possible Futures, Affective Information Processing (Tao, J., Tan, T., eds.), pp. 111-126, 2009.

[4] Schröder, M., Approaches to emotional expressivity in synthetic speech. In: K. Izdebski (Ed.) The emotion in the human voice, vol 3, Plural, San Diego, 2008.

[5] Morrison, D., Wang, R., & De Silva, L. C., Ensemble methods for spoken emotion recognition in call-centres. Speech Communication, 49, pp. 98–112, 2007.

[6] Murray, I. R., & Arnott, J. L., Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustic Society of America, 93(2), pp. 1097–1108, 1993.

[7] Burkhardt, F., & Sendlmeier, W.F., Verification of acoustical correlates of emotional speech using formant synthesis. In Proc. of the ISCA Workshop on Speech and Emotion, Northern Ireland, pp. 151–156, 2000.

[8] Fernandez, R., & Ramabhadran, B., Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 34–39, 2007.

[9] Zen, H., & Toda, T., An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In Proc. of InterSpeech, Lisbon, Portugal, pp. 93–96, 2005.

[10] Campbell, N., Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. In Proc. of the Intl. Congress of Phonetic Sciences, Saarbrücken, Germany, pp. 343–348, 2007.

[11] http://www.w3.org/TR/voicexml20/

[12] http://www.w3.org/TR/speech-synthesis/

[13] http://www.w3.org/TR/jsml/

[14] http://www.bell-labs.com/project/tts/sable.html

[15] http://mary.dfki.de/documentation/maryxml

[16] Wiebe, J., Wilson, T., and Cardie, C., Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3):165-210, 2005.

[17] Pang, B. and Lee, L., Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proc. 43rd Annual Meeting of the ACL, pp. 115-124, Michigan, 2005.

[18] Shaikh, M. A. M., Prendinger, H., and Ishizuka, M., Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. Applied Artificial Intelligence, Vol.22, Issue 6, pp.558-601, Taylor & Francis, 2008.

[19] Liu, H., Lieberman, H., and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international Conference on intelligent User interfaces* (Miami, Florida, USA, January 12-15, 2003). IUI '03. ACM, New York, NY, 125-132.

[20] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," Intl. J. Speech Technol., vol. 6, pp. 365–377, 2003.

[21] Machinese Syntax http://www.connexor.eu/technology/machinese/

[22] A. Ortony, G. L. Clore, and A. Collins, The Cognitive Structure of Emotions. Cambridge University Press, July 1988.